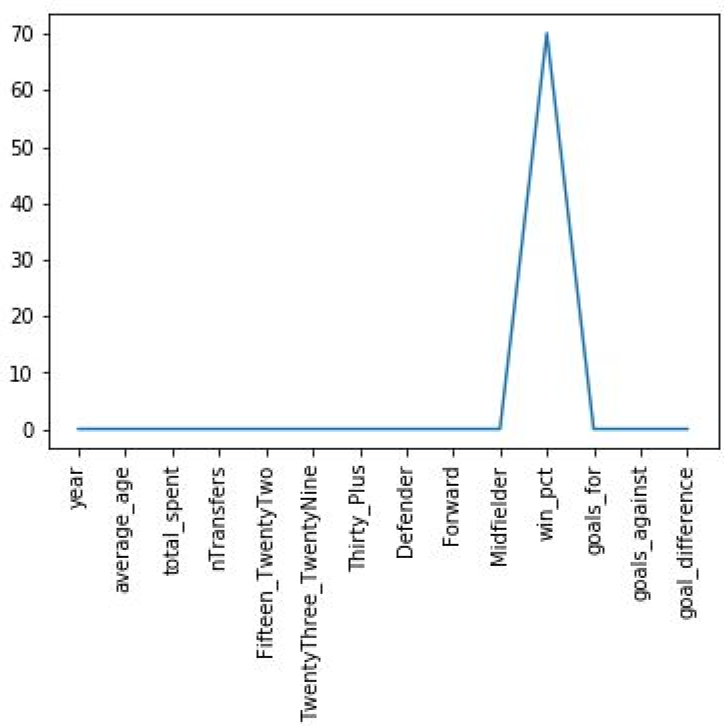
In-Depth Analysis

I applied two different machine learning techniques to my data. I first wanted to test for the most important independent variable I had collected, so I chose to use a Lasso regression, so as to shrink all of the unimportant variables to zero. I first split my data into train and test subsets with the sklearn.model_selection tool train_test_split. Then, I made a Lasso model, fit the training set to my data, and checked the coefficients and score for the test set. The only variable that was not shrunk to zero in this model was win percentage, which had a Lasso coefficient of 70.07. The test set had a score of 0.598, so about 60% of the variance in this model can be explained by win percentage. The root mean squared error was about 13.5, which is high, leading me to believe it is more than just the win percentage that can explain a team's position value.

Lasso Regression Coefficients



I also implemented random forests in my data, so as to see the importances of my features, but not shrink all but one of them to zero. I created a random forest regressor with the sklearn.ensemble package and fit the model to the training set. Obtaining the score on the test set yielded a value of 0.857, which is significantly higher than the Lasso regression. The root mean squared error was also lower at 8.09. The most important feature was still win percentage at 0.64, but goal difference (goals scored minus goals allowed) could also be considered important at 0.24. All of the teams transfer data showed low scores on these models, leading me to conclude that a team's game performances are generally not affected by their spending.